Chapter 7 Field Test Design, Sampling, and Administration

Introduction

A major goal of the Field Test Administration was to provide validity evidence in support of the Smarter Balanced summative and interim assessment purposes. The final Smarter Balanced scales and supporting elements were established in the Field Test using Smarter Balanced Consortium schools, districts, and states that were engaged in the process of implementing the Common Core State Standards (CCSS). The design, while complex, was efficient both in the testing time projected and in the number of items necessary to meet the program requirements. The test design called for each student to be exposed to the full test blueprint and mix of item types that included the less familiar performance task (PT) component. A targeted "Standard Setting Sample" was used to establish the final horizontal and vertical scales and provide the information used to set achievement levels. In a second step, a larger item-pool calibration sample was used for scaling and horizontally linking a robust set of items onto the Smarter Balanced scale established in the previous step. This second calibration step represents the entire item pool at the conclusion of the Field Test. More detail concerning the scaling and linking designs can be found in Chapter 9 Field Test IRT Scaling and Linking Analyses. The items intended for the operational CAT (Computer Adaptive Testing) administration were delivered using Linear-on-the-Fly-Testing administrations. This was advantageous since a content-balanced test blueprint can be delivered to each student using a delivery mode closer to the operational CAT. To the extent possible, samples were selected to represent the demographic characteristics of the Smarter Balanced Governing States. The primary purpose of this chapter is to describe the purposes, design principles, and implementation requirements for the Smarter Balanced Assessment Consortium Field Test administration and its results.

Smarter Balanced conducted a Pilot Test administration in 2013 to inform some aspects of the Smarter Balanced assessments. The Pilot further informed the item types to retain or the ones with entirely new formats to develop as well as the revisions to the test blueprints. Essential elements of the design were changed, such as the inclusion of Classroom Activities in the PTs. Another outcome from the Pilot was the selection of the Item Response Theory (IRT) scaling models. Based on the Pilot analysis, the unidimensional two-parameter model and the generalized partial-credit model for mixed-format items were chosen. While some expected Pilot Test outcomes needed to be revisited in the Field Test, there were somewhat different goals targeted for the Field Test. The major purposes of the Field Test administration were

- to administer and calibrate a sufficient large number of items to ensure a successful operational launch of summative and interim assessments;
- to obtain classical statistics and produce Differential Item Functioning (DIF) analyses to inform item data reviews;
- to establish the final operational horizontal and vertical scales;
- to set the achievement level standards;
- to evaluate the protocols for the test administration and computer delivery system (technology infrastructure); and
- to implement targeted test accommodations and elements of universal design.

The Field Test administration window extended from March 18 to June 6, 2014, for all participating states. In order to achieve these varied purposes, 15,673 items resulted from the Field Test for ELA/literacy (ELA) and mathematics across all grade levels. These items and tasks were delivered to 1,742,208 students from the Smarter Balanced Governing States. To support IRT calibrations, the

number of responses for each item was targeted at 1,200 observations. In many instances, items with fewer than 1,200 observations were calibrated if 500 cases were available. The student samples will be drawn from Smarter Balanced Governing States according to the same two stage, sampling strategy used for the Pilot Test. Some additional items for special study were also included, from the National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA) items at selected grades. These items were placed onto the Smarter Balanced vertical scale and were used in the achievement level setting (standard setting) in the fall of 2014. External items from NAEP and PISA were linked onto the Smarter Balanced scale horizontally using on-grade common items. The relative difficulty of these items was compared with Smarter Balanced items to obtain external measures of performance that could inform the achievement-level setting process.

Field Test Data Collection Design

There were two overall basic steps to the Field Test Design. The first step in the analysis was to establish the vertical and horizontal scales using a robust sample. Items were designated either on-grade or off-grade for vertical linking purposes. Vertical linking items were given across two grade levels (i.e., common items) using content from the lower adjacent grade (e.g., fifth-grade items given to sixth grade). Each student also took a performance task in order to conform to the test blueprint that could be on-grade or off-grade. These items and samples were also used in the achievement-level setting. A representative sample of Smarter Balanced test content and students were needed in order to construct the ordered-item booklets and impact data for the achievement-level setting. The external items from NAEP and PISA were also targeted at the standard settings that were given in selected grades. PISA items were administered in grade 10 for Smarter Balanced. NAEP was given in grades 4, 8, and 11 (in lieu of grade 12). The second step was used to calibrate large numbers of items horizontally in a grade to populate the main IRT item pool. All items administered in the vertical linking step were also administered on-grade to the calibration sample and served as the “common items” for linking. The calibration sample was then linked back to the vertical scale established in the first step using the common/“anchor” items in a grade. The CAT items were administered using Linear-on-the-Fly-Testing, while the performance tasks were fixed forms (not computer adaptive) administered online. The vertical scaling and item pool calibration were separate student samples.

To administer items for vertical scaling, four delivery conditions (summarized in Table 1) were employed in the data collection.

- Condition 1: Tests delivered in this condition include approximately 50 content-representative CAT items and an on-grade or off-grade performance task.
- Condition 2: Tests administered here include approximately 25 content-representative on-grade CAT items, approximately 25 content-representative upper grade CAT items, and an on-grade or off-grade performance task.
- Condition 3: Tests delivered under this condition include approximately 25 content-representative on-grade CAT items, approximately 25 content-representative lower-grade CAT items, and an on-grade or off-grade performance task.
- Condition 4: In this condition, approximately 25 content-representative on-grade CAT items, approximately 25 content-representative NAEP or PISA items, and an on-grade or off-grade performance task were included.

Condition 1 was used to calibrate a large number of items on-grade. Conditions 2 and 3 were targeted at the vertical scaling. Condition 4 was used to calibrate NAEP and PISA items onto the Smarter Balanced scale.

Table 1. Field Test Data Collection Design for ELA/literacy and Mathematics.

Condition	CAT Items	CAT Assignment	P T Assignment	External Items
Vertical Scaling Step				
1	~50 CAT	On-grade only	1 on/off grade	
2	~50 CAT	On-grade/Upper Grade	1 on/off grade	
3	~50 Cat	On-grade/Lower Grade	1 on/off grade	
4	~25 CAT	On-grade only	1 on/off grade	25NAEP/PISA
Calibration Step				
Item Pool Calibration	~50 CAT	On-grade only	1 on-grade	

Field Test Design Principles. The following design principles underpinned the data collected to ensure the best outcomes for the item analysis, calibration, and construction of the vertical scales.

- The tests presented to (and scored for) each student conform to specified test blueprint content requirements. Analyses and resulting scores are more interpretable when the test form administered to each student is appropriately and consistently content balanced. The analyses assumed that students were presented interchangeable test forms measuring essentially the same construct.
- The second principle concerned the linking for the vertical scale being based on substantial item collections administered to representative student samples across grades.
- Thirdly, items should be administered at approximately uniform rates. Assembly specifications for Linear-on-the-Fly-Testing should be detailed and firm enough to ensure consistency of the trait(s) measured but not so rigid that they force distinctly unbalanced rates of item use. Finally, items should be administered to substantial student samples.

In addition, the Field Test had the following characteristics and specifications. All content strata and item types were available to represent the construct in a grade and were administered throughout the entire testing window. The design incorporated two overlapping item and student samples, which consisted of the CAT and performance tasks that were separately delivered events. There was no distinction between summative and interim item pools in these calibration steps. After the calibration was completed and all IRT statistics were available, the items were partitioned into the summative and interim pools. Psychometric characteristics such as item response time, item exposure rates, and specifications for CAT algorithms were not examined due to the information not being available or occurred in other later phases of the program.

Field Test Delivery Modes

For the Field Test, the test delivery modes corresponded to the two separately delivered events. The performance tasks were delivered using computerized fixed forms/linear administrations. For a given performance task, students saw the same items in the same order of presentation and

associated test length. Since performance tasks had a classroom-based activity and were organized thematically, they were randomly assigned at the school level in the Field Test.

CAT (LOFT) Administration. For the CAT pool in the Field Test, Linear-on-the-fly testing (LOFT) was used to administer items to students (Gibson & Weiner, 1998; Folk & Smith, 2002). Note that the LOFT is similar to a CAT in applying content constraints to fulfill the test blueprint. LOFT delivered tests that were assembled dynamically to obtain a unique test for each student from a defined item pool where each student obtains a unique content-conforming test form. The major differences between LOFT and item- level adaptive testing are that no IRT item statistics are used in the administration, and no adaptation based on student responding/ability is incorporated into the delivery algorithm. For dynamic real-time LOFT, item exposure control (e.g., Hetter & Sympson, 1985) can be used to ensure that uniform rates of item administration are achieved. That is, it is not desirable to have some items with many observations and others with correspondingly few in comparison. The LOFT administration is closer to the operational CAT than fixed forms. This permits the scaling to reflect the operational CAT deployment. The major advantage of using LOFT was that delivering parallel fixed test forms with thousands of items in a pool in a given grade and content area was not possible. The disadvantage is that some measures of test functioning are not directly available using LOFT. Observed score (i.e., classical) statistics such as observed test reliability cannot be computed since every student essentially takes a unique test form. Even the definition of a criterion for item-test correlation and for DIF must rely on IRT methods for computing these statistics.

Performance Task Administration. In the case of performance tasks, a Classroom Activity was assigned by school and grade. Four to six separate performance tasks were associated with each Classroom Activity and were to be spiraled to all students at a grade level within a school. Smarter Balanced item and task specifications assumed computer delivery of the items and tasks. Most tasks were long enough to warrant several administration sessions. Such sessions could be same-day, back-to-back sessions with short breaks between sessions. All tasks were administered in controlled classroom settings. Expected time requirements for completing tasks and administration time were provided in subject-specific specifications. Student directions for all tasks began with an overview of the entire task, briefly describing the necessary steps. The overview gives students advanced knowledge of the scorable products or performances to be created (Khattri, Reeve & Kane, 1998). Allowable teacher-student interactions for a task were standardized (i.e., carefully scripted or described in task directions for purposes of comparability, fairness, and security). Teachers were directed not to assist students in the production of their scorable products or presentations.

The group work and teacher directions on how to form and monitor groups for the classroom component of the PTs ensured that no students are disadvantaged simply because of the group to which they are assigned. Group work was not scored but was designed to accomplish such things as the generation of data, the discussion and sharing of provided information, or role-playing for the purposes of the task. If small-group discussions could potentially advantage some groups, the teacher directions required them to use standardized scripts to summarize key points that should have come out of the group discussions. Procedures for standardizing the group-work component will vary depending on the task type. Some task steps will require teachers to play more than a monitoring role and/or students to do small-group work. Teachers and peers were directed not to assist students as they produced their scorable products. The permitted types of teacher and peer interactions for a task were standardized (i.e., carefully scripted and explicitly described in task directions) for purposes of both fairness and security. Although small-group work may be involved in some part of a task, this part was not scored. Students were informed about the nature of the final product(s) at the beginning of the task. The task directions included information for the students on what parts of their work would be scored. All scorable products or performances reflected individual

student products. Every task had multiple scorable products or performances. With responses such as essays, students were informed about which attributes of their work would be scored.

Field Test Design

To achieve the larger Smarter Balanced goals for summative and interim operational assessments and conduct the Field Test, both items and accompanying student samples were targeted to fulfill the scaling and calibration requirements. Table 2 indicates the number of CAT items and performance tasks targeted to support both summative and interim test purposes. To calculate the number of tasks, the assumptions were that each ELA/literacy task would have four items (i.e., scoreable units) and each mathematics performance task would contain six items. Some important scaling-design decisions entailed in Table 2 are listed below. For example, 1,290 items in total were targeted for administration in grade 3. Approximately 300 items were delivered to the vertical scaling sample and 990 items to the item-pool calibration sample. Similar sorts of things pertained to the performance tasks—53 items were targeted for development in total, with 47 in the item pool calibration and 6 administered in the vertical scaling step.

- Students were assigned either ELA/literacy or mathematics in order to minimize the burden of testing time on schools.
- The number of CAT items necessary was estimated using the ratio of a test blueprint (approximately 50 items) to the entire pool as 10:1, which is consistent with judgments for CAT pool size. This was used as a rule-of-thumb to project the number of CAT items needed in a grade/content area to support the Field Test purposes. For example if the blueprint required 50 items for an individual test administration, 500 items collectively would then be needed using this rule.
- The number of performance tasks was determined by their anticipated exposure rates and attrition from the summative pool. Additional items were developed for ELA/literacy due to reading passages and listening in which items are clustered and more flexibility is desired in item selection.
- To achieve these numbers in the operational tests, a 10% overage for CAT item development and a 20% overage for performance tasks development were used to account for expected item attrition during development. For example, in the case of a target of 300 CAT items, 330 were developed but 10% might not be expected to survive content or bias and sensitivity reviews.
- Under this plan, fewer than half the items were targeted for the interim tests compared with the summative tests, with approximately one-third the number of performance tasks used for the interim tests.
- Items were written for grade 11 using the Common Core State Standards were also administered at grades 9 and 10 for vertical scaling.
- The number of CAT items in the interim system was estimated to be 50% of those in the summative system. The number of performance tasks in the interim system was targeted to be 25% of those contained in the summative system.
- In operational settings, the interim pool will be “refreshed” using items that are retired from the summative tests.
- These numbers reflect grade-specific deployment. With a vertical scale, items are used across several grades in order to perform the linking. In this case, items were administered

to the upper-adjacent grade (e.g., grade 5 items given to grade 6 students), expanding the number of available items in a grade.

- Each item/task will have at least 1,200 valid student responses entering the analyses, assuming that uniform exposure control and item pools are proportional to LOFT blueprints. A sample of 1,200 observations was sufficient to obtain reasonably accurate statistics for the 2PL and GPCM IRT scaling models (Stone & Zhang, 2003). A minimum of 500 cases was necessary for inclusion in the calibrations.
- All items in the vertical scaling pool were also administered on-grade in the item-pool calibration step for IRT horizontal linking purposes.

Table 2. Field Test Design for the Item and Performance Task Pools.

Grade	Total		Vertical Scaling		Item Calibration Pool	
	CAT	PT	CAT	PT	CAT	PT
ELA/literacy						
3	1,290	53	300	6	990	47
4	1,290	53	300	6	990	47
5	1,290	53	300	6	990	47
6	1,290	53	300	6	990	47
7	1,290	53	300	6	990	47
8	1,290	53	300	6	990	47
HS* (9,10,11)	3,765	144	300	6	3,465	138
Mathematics						
3	1,125	54	300	6	825	48
4	1,125	54	300	6	825	48
5	1,125	54	300	6	825	48
6	1,125	54	300	6	825	48
7	1,125	54	300	6	825	48
8	1,125	54	300	6	825	48
HS (9,10,11)	3,435	150	300	6	3,135	144

**Note: HS refers to High School.*

Numbers and Characteristics of Items and Students Obtained in the Field Test

The sample size for the Field Test is determined by the total number of items to be field tested, the sample size required for each item, and the specific field-testing design. A targeted sample size of 1,200 valid cases for each item was needed to support the production of classical statistics and IRT calibrations. This placed a premium on the item-exposure rates being relatively uniformly distributed since an established sample size was targeted.

When an item or a task was used off-grade for vertical scaling, the effective number of observations for the item/task roughly doubled. Items given on-grade for the standard setting sample were administered as common items in the calibration sample, effectively doubling the observations collectively for item collections. In addition, some statistics will require designated samples, such as students with disabilities and English Language Learners (ELLs). The sizes for the special samples will need to permit differential item functioning (DIF) comparative analysis. The Mantel-Haenszel (MH) and Standardized Mean Difference (SMD) procedures were implemented for DIF studies with IRT ability (theta) as the matching criterion. The minimal sample size for the focal or reference group is 100 and for the total (focus plus reference) group is 400.

Tables 3 and 4 show summaries of the subset of items and tasks that were used for vertical scaling and NAEP/PISA that resulted after test delivery and item exclusions. The distribution of the items according to the claims is also presented. In most cases, the number of linking items was robust. In high school mathematics, there was additional attrition of vertical linking items. A smaller set of items was subsequently eliminated in the IRT scaling step not reflected here. Tables 5 and 6 show the items obtained after the item pool calibration step by claim.

Table 3. Number of Field Test Vertical Scaling Items Obtained by Type for ELA/literacy.

	Grade						
	3	4	5	6	7	8	HS
	ELA/literacy						
Total	261	362	389	363	345	366	517
Claim 1	94	112	145	124	116	132	229
Claim 2	70	101	99	98	99	100	151
Claim 3	50	77	71	70	64	73	60
Claim 4	47	72	74	71	66	61	77
Claim Unknown							
NAEP		28				30	27
PISA							33
Off-grade		120	133	131	107	123	107
On-grade	261	242	256	232	238	243	410

Table 4. Number of Field Test Vertical Scaling Items Obtained by Type for Mathematics.

	Grade						
	3	4	5	6	7	8	HS
	Mathematics						
Total	304	440	401	324	310	336	502
Claim 1	184	240	237	167	162	166	237
Claim 2	17	21	20	24	14	18	27
Claim 3	47	67	67	53	53	44	56
Claim 4	19	30	26	26	27	24	39
Unclassified	37	82	51	54	54	84	143
NAEP		30				33	28
PISA							74
Off-grade		104	95	102	71	73	81
On-grade	304	306	306	222	239	230	319

Table 5. Number of Field Test Item Pool Calibration (Overall Total) Obtained by Claim for ELA/literacy.

	Grade						
	3	4	5	6	7	8	HS
	ELA/literacy						
Total	896	856	823	849	875	836	2,371
Claim 1	317	259	265	274	299	258	867
Claim 2	243	248	241	257	262	241	729
Claim 3	163	157	142	147	152	174	383
Claim 4	173	192	175	171	162	163	392

Table 6. Number of Field Test Item Pool Calibration (Overall Total) Obtained by Claim for Mathematics.

	Grade						
	3	4	5	6	7	8	HS
	Mathematics						
Total	1,114	1,130	1,043	1,018	942	894	2,026
Claim 1	672	677	613	576	519	493	1123
Claim 2	55	68	55	77	71	59	147
Claim 3	166	145	168	132	120	134	433
Claim 4	68	77	84	68	67	64	185
Unclassified	153	163	123	165	165	144	138

Figures 1 and 2 on the following pages show the frequency distributions for the number of items delivered to students (i.e., test length) for ELA/literacy and mathematics for the vertical scaling. Grades are shown both together and individually either in ELA/literacy or mathematics. These item

counts per student included both the CAT and PT components. In ELA/literacy, bimodal distributions in test length are evident for all grade levels. Many students received approximately a 25-item CAT along with a PT. This was due in part to the inclusion of California, where two half-tests were administered, being resampled to maximize item exposure rates in the delivery system. In mathematics grades 6, 7, and 8, the resulting test length was comparatively short and bimodal only in selected grades. Figures 3 and 4 show similar sorts of information for the item pool calibration.

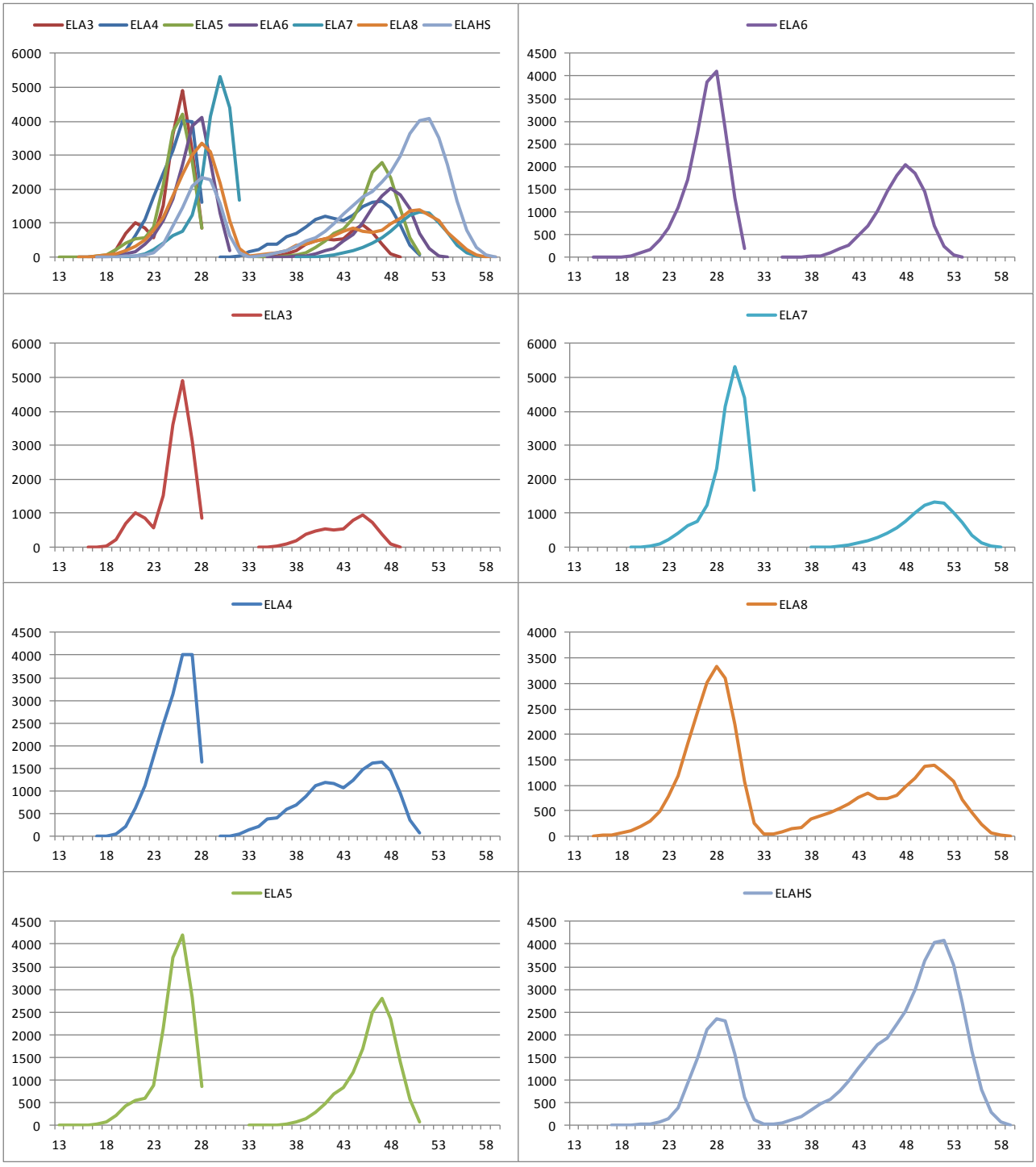

Figure 1. Distributions of Number of Items per Student in the Vertical Scaling (ELA/literacy)

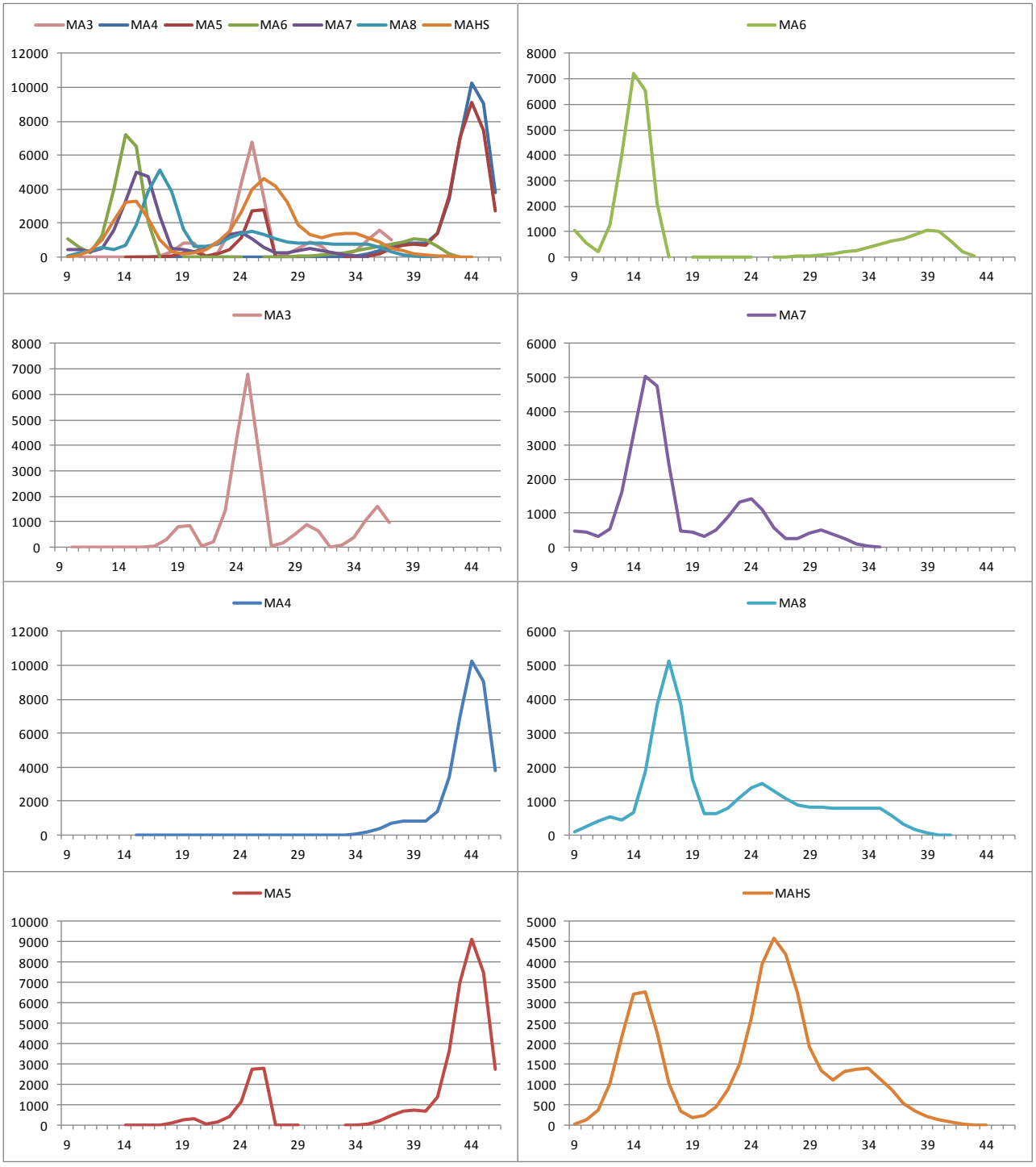

Figure 2. Distributions of Number of Items per Student in the Vertical Scaling (Mathematics)

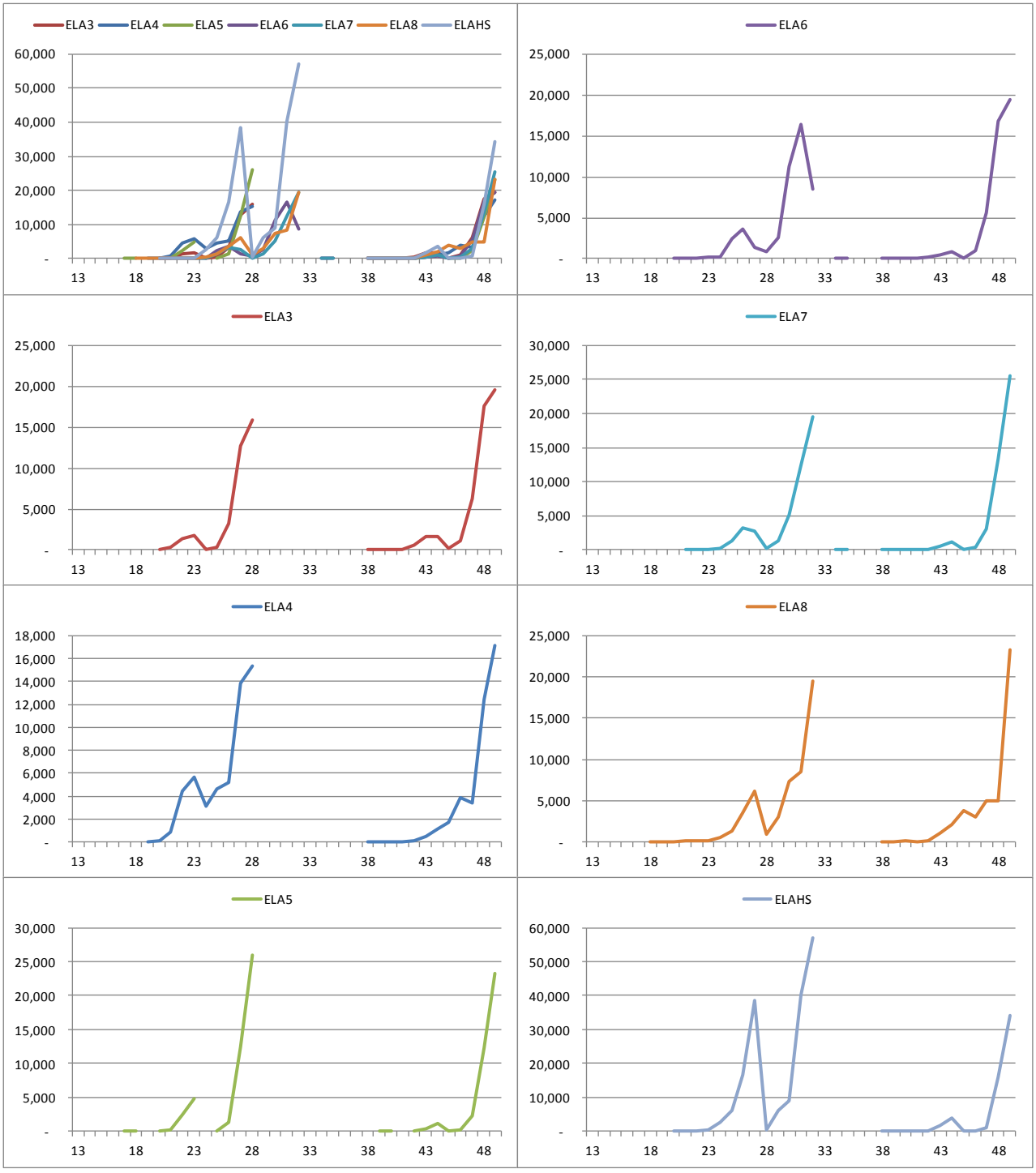

Figure 3. Distributions of Number of Items per Student in the Item Pool Calibration Sample (ELA/literacy)

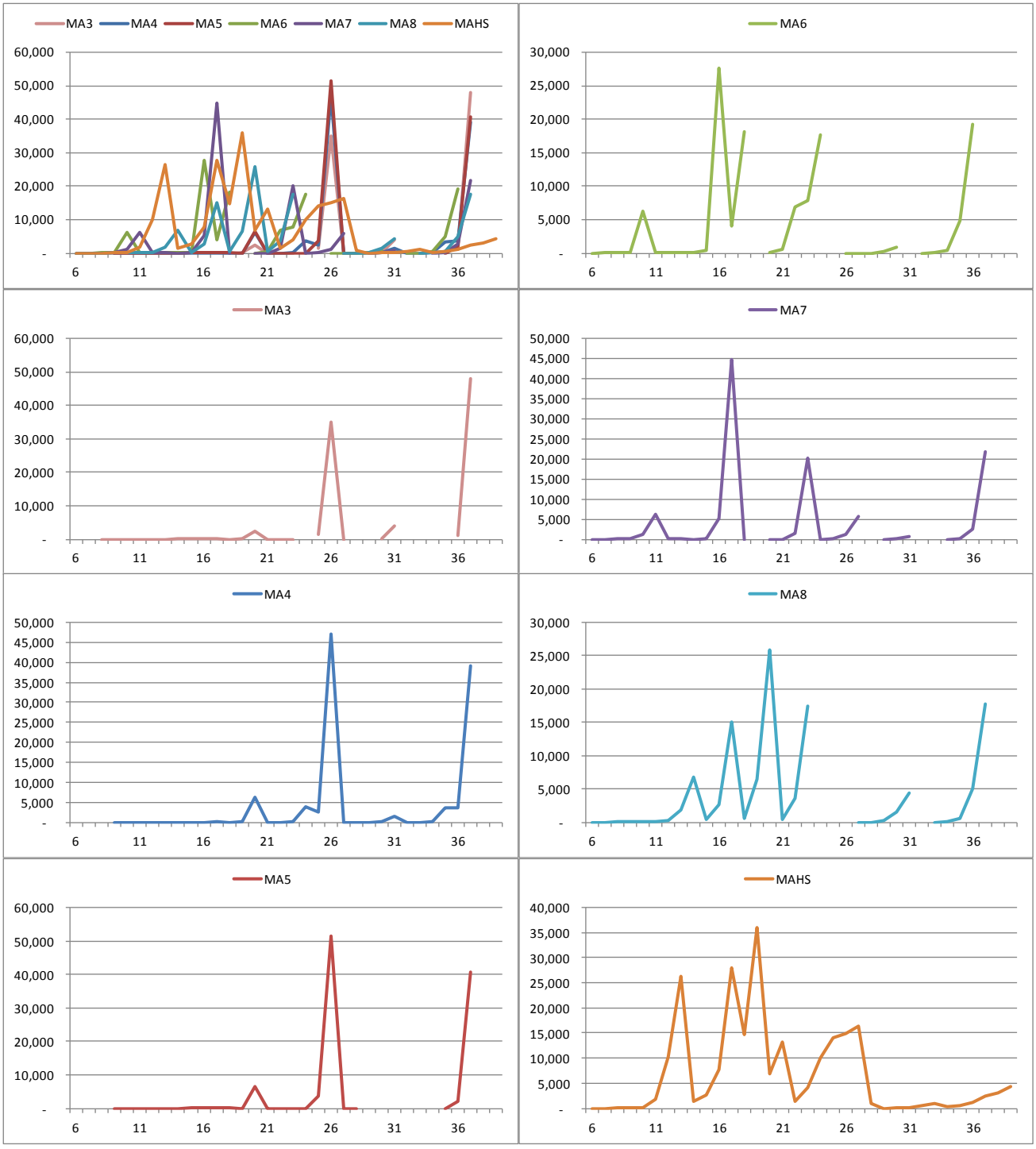

Figure 4. Distributions of Number of Items per Student in the Item Pool Calibration Sample (Mathematics)

Linking PISA and NAEP Items onto the Smarter Balanced Assessments

In the Smarter Balanced Theory of Action, a goal was to establish clear, internationally benchmarked performance expectations. To inform achievement-level setting for Smarter Balanced inferences concerning national and international performance, item collections were obtained from the NAEP and PISA programs. In the United States, national-level data on student achievement stems primarily from two sources: the National Assessment of Educational Progress (NAEP)—also known as the "Nation's Report Card"—and participation in the Program for International Student Assessment (PISA). NAEP measures fourth-, eighth-, and twelfth-grade students' performances, most frequently in reading, mathematics, and science, with assessments designed specifically for national and state information needs. Alternatively, the international assessments allow the United States to benchmark its performance to that of other countries in 15-year-olds' reading, mathematical, and scientific literacy with PISA. These assessments are conducted regularly to allow the monitoring of student outcomes over time. While these assessments appear to have some general similarities, such as the age or grade of students or content areas studied, each program was designed to serve a different purpose and each is based on a separate and unique content framework and set of items. The major features of Smarter Balanced, NAEP, and PISA assessment programs are compared in Table 7.

The Field Test provided the initial opportunity to link selected external items onto Smarter Balanced assessments. A special Field Test data collection condition was required to support this goal. In a secondary step after the vertical scaling calibration of Smarter Balanced items, these external items were calibrated and linked to obtain IRT item parameters on Smarter Balanced scales. This required a content-representative collection of Smarter Balanced Field Test items in ELA/literacy or mathematics and a collection of NAEP and PISA items to be administered to designated students. These external items replaced the off-grade CAT items. Smarter Balanced used released PISA and NAEP items for this purpose. NAEP items were embedded in Smarter Balanced grades 4, 8, and 11 assessments in both content areas. After calibration of Smarter Balanced items, PISA and NAEP were calibrated onto Smarter Balanced scale(s) using randomly equivalent samples and common items. Recognizing these differences in the nature of the construct and test purposes between these programs, the resulting item parameters on the Smarter Balanced scale were used to inform inferences concerning relative performance in the Smarter Balanced achievement-level setting.

Table 7. Comparison of Features across the Smarter Balanced, NAEP, and PISA Assessment Programs.

Design Feature	Smarter Balanced	NAEP	PISA
Construct Definition	ELA/literacy Claims—Reading, Writing, Listening, & Research Text Types: Literary & Information	Reading Frameworks—(writing is separate) Text Types: Literary & Information	Reading Aspects— Text Types: Exposition, Argumentation Instruction, Transaction, & Description
	Math Claims—Concepts and Procedures, Problem solving, Model and Data Analysis, Communicating Reasoning	Math Frameworks— Number Properties and Operations, Measurement, Geometry, Data Analysis, Statistics and Probability, and Algebra	Math Aspects— Quantity, Uncertainty, Space & Shape, Change & Relationships
Item Context Effects and Test Administration Rules	The look and feel of NAEP and PISA items will likely be different from Smarter Balanced items. The provision of glossaries, other test manipulatives, and accommodation rules will differ across programs. Smarter Balanced uses technology-enhanced items, while PISA and NAEP do not.	The basic context will be maintained for NAEP items since they are administered as a set(s).	The basic context will be maintained for PISA items since they are administered as a set(s).
Testing Mode	LOFT delivery on computer and PTs online	Paper 2015: paper scale and computer-based testing scale study	Paper 2015: computer-based testing scale
Testing Window	March–June 2014	February 2013	PISA in April/May
Untimed/Timed	Untimed	Timed	Timed
Delivery Design	Smarter Balanced Field Test LOFT blueprint(s) that took into consideration the embedded set(s) properties, such as their testing length, reading load, and associated number of items	Linear Administration	Linear Administration
Constructed-Response Scoring		Approximately 30–40 % of the NAEP items require rater scoring. Scoring protocols such as training and	Approximately 30% of the PISA items associated with set(s) require rater scoring. Scoring protocols such as

Design Feature	Smarter Balanced	NAEP	PISA
		qualification will need to be followed. Handwritten responses would need to be transcribed for anchors, training and qualification, and calibration papers.	training and qualification will need to be followed. Handwritten responses would need to be transcribed for anchors, training and qualification, and calibration papers.
Cohort/ Population	Sample of 2014 Smarter Balanced Governing States	Based on 2013 US national sample with state-level comparisons	Based on 2012 US sample: 5,000 15-year-old students from 150 schools
Criterion-Referenced Inferences	Designated achievement-level scores in 2014	Proficiency cut scores exist.	Proficiency cut scores do not exist.
Anticipated Program Changes	No change after 2014 in content; schools still transitioning to the CCSS	Transitioning to computer based administration in 2015	Computer based in 2015 and assessment framework will change
IRT Model and Scaling Procedures	Scaling is at the overall content area level using the tow-parameter logistic (2-PL)/ generalized partial credit model (GPCM).	3-PL and GPCM in reading and math: The main scales are weighted composites of subscales, and calibration is done at the subscale level.	Rasch (calibrated separately with relation to major domain and minor domain)

Field Test Student Sampling Design

Given the purposes and the nature for Smarter Balanced assessments, it is important that the resulting test scales and associated achievement levels represent the performance characteristics of the participating Smarter Balanced states. The characteristics of the Field Test sample will ultimately be reflected in the item statistics and the scales developed.

The Field Test study targeted a representative sample as opposed to a convenience sample of volunteering (self-selected) schools. A multiple-stage stratified sampling with nested cluster sampling was used as the primary approach. The selected samples were intended to be representative of the intended population, which consists of all students from Smarter Balanced member states. The same sampling procedure will be used to recruit samples for nine grades (3–11) and two content areas (mathematics and ELA/literacy), totaling 18 separate samples. The Field Test mirrored the operational test to the extent that every student was designated to take both a CAT and a PT component. An exception was in California where students took both ELA/literacy and mathematics (half-length tests) and a single performance task in the item pool calibration. In the context of vertical scaling and the standard setting sample, some students were assigned both off-grade and on-grade test content configurations. Some states elected to use their own procedures to select representative student samples for the Smarter Balanced Field Test. A volunteer sample was also

collected but was not included in the formal sampling and test analysis. There was no oversampling of any particular subgroup.

Using the test designs given in Table 1, three different Field Test conditions and associated student samples were used that corresponded to assignment to the vertical scaling, item pool calibration, and volunteer conditions. Using school demographic characteristics, such as ethnicity or percentage proficient, a representative sample was selected separately in each state for the vertical scaling and calibration samples. The volunteer schools were used as replacement schools when necessary.

1. The vertical scaling sample (First Sampling Priority) took a content-representative sample of CAT items and performance tasks sufficient to implement vertical and horizontal scaling and construct ordered-item booklets for standard setting. It was essential that the CAT items and their content characteristics closely follow the desired operational pool and the test blueprints. For vertical linking items, there were additional students at two grade levels. Grades 9, 10, and 11 students included under "high school" were used for vertical linking of grades 8 to 11. A subset of students took NAEP/PISA items sufficient for scaling purposes. The sample size targeted for these items was the same as the Smarter Balanced ones. The items in this pool were targeted for delivery to a representative sample from participating Smarter Balanced Governing States. Since this group was used to determine the Smarter Balanced vertical scale, considerable effort was directed at identifying obtaining a representative sample.
2. Calibration Sample (Second Sampling Priority) consisted of students taking all items and tasks on-grade level. The goal was to calibrate a very large number of items in the remaining pool. This pool also included common items from the high-priority vertical scaling pool used to link them onto the final scale. Administration of vertical linking items was not necessary here since this was accomplished in the vertical scaling condition.
3. Volunteer Sample. The remaining participating students were volunteers. More students participated in the Field Test than were needed for scoring and scaling. Since the characteristics of these schools were known, they could be used as replacement schools for the vertical scaling and calibration sample when necessary. These students took both CAT items and a PT, and they may have been required to take tests in one or both content areas.

Defining the Target Population. The defining of the target population provides characteristics for evaluating the representativeness of the resulting sample and the sampling strategies used to obtain it. There were several factors considered in defining the characteristics of the target population for the Field Test, including the model for representing state participation, transition to the Common Core State Standards, and technology infrastructure available for testing.

To be representative of the target population, Field Test samples were recruited to have state representation that was proportional to the size of the state's student enrollment ("House of Representatives" model). The percentages constitute an implicit sampling weight for each state that is reflected in the vertical scaling item- pool calibration samples. The other model not adopted was the "Senate," where an equal number of students are contributed by each state. Per Smarter Balanced recommendations, Advisory and Governing States both participated in the Field Test where proportional representation was implemented without considering a state's "governing" or "advisory" status. A two-state stratified random sampling was used in each member Governing State; the first stage was the state, with schools within the state as the second stage. States were sampled in proportion to their student enrollment size. It was not possible to completely control for situations where states either dropped out of the Consortium or were added after the Field Test.

The second factor considered in defining the target population is the level of Common Core State Standards implementation. Among Smarter Balanced states, the extent to which the CCSS was

implemented at the time of the Field Test administration was likely to vary considerably and no accurate information was available. The target population consisted of students from all Smarter Balanced member states and schools, regardless of the Common Core State Standards implementation level. It is likely that some scale drift will occur over time as the Common Core State Standards are more fully implemented.

The final factor considered in the definition of the target population is the capacity for online testing. While some states are currently administering online state assessments, other states may have districts and schools with varying capacity for online testing. Schools needed to have the specified level of technology infrastructure in order to participate in the Field Test. The necessary technology specifications were communicated to schools. The selected samples include students from schools with varying capacity for online testing. To some extent, the level of technology infrastructure drove the decision-making concerning the number of students that can be selected and reasonably tested online in a given school.

State Participation Conditions. Smarter Balanced states used four types of state participation models for the Field Test data collection. In August 2013, states were asked to provide their anticipated participation model. Table 8 shows the states that were initially planned to be part of the Smarter Balanced 2014 Field Test, as well as whether they recruited/selected their own sample. States could either work with the Smarter Balanced test administration workgroup or implement their own sampling that had to adhere to established criteria for representativeness. Note that some of the state participating models may have changed status due to waiver requests and other state policy decisions. The state participation models were the following:

- Early Adopter states required full participation in both content areas of the Smarter Balanced Field Test in the 2013-14 school year in place of the state's accountability test.
- Blended Basic states constitute states that committed to the number of schools minimally necessary to fulfill the prescribed Field Test sample. These schools did not take the state's accountability test in the 2013-14 school year.
- Blended Enhanced states are states that committed to more than the prescribed Field Test sample to participate in the Field Test (i.e., allowing more students than the prescribed sample but less than 100%). These schools do not take the state's accountability test in the 2013-14 school year.
- Traditional states are ones that require all schools to administer the existing state's accountability test (traditional administration) and committed schools to participate in the Smarter Balanced Field Test in the 2013-14 school year minimally necessary to fulfill the state's prescribed Field Test sample.
- Affiliate or Advisory States are any Smarter Balanced Affiliate or Advisory State member that elects to participate in the Field Test and participated in a strictly voluntary mode.

Throughout the recruiting cycle (September to February), state-participation-status changes occurred, such as Kansas withdrawing from the Consortium, Wisconsin opting not to test high school, North Carolina opting not to require field testing, California choosing to be a modified Early Adopter state, and Missouri opting not to recruit for high school. Adjustments were made to the sample where possible by proportionally allocating these cases to other states.

Table 8. State Participation and Sample Acquisition Conditions.

Condition	State-Led Sampling	Smarter Balanced-Led Sampling
Early Adopter	South Dakota	Idaho Montana
Blended Basic	Nevada Vermont	Kansas Michigan Oregon
Blended Enhanced	Washington Connecticut	California
Traditional	Missouri North Carolina West Virginia Wyoming (Plus) Iowa	Delaware Hawaii (Plus) Maine (Plus) New Hampshire North Dakota South Carolina Wisconsin (no HS)
Affiliate/Advisory	Virgin Islands (VS)	

Technical Sampling Characteristics. In the Smarter Balanced sampling design, the impact of different sampling conditions or procedures was considered. The sampling factors considered were the smallest unit of sampling, the use of sample weights, nonresponse, and sampling from voluntary districts/schools. These are discussed below.

- **Smallest sampling unit.** Whereas stratification generally increases precision when compared with simple random sampling, cluster sampling generally decreases precision. Simple random sampling at the student level cannot be conducted in educational settings since students usually reside within classrooms. In practice, cluster sampling is often used out of convenience or for other considerations. If clusters have to be used, it is usually desirable to have small clusters instead of large ones. The recommendation is to have an entire grade level from a school as the smallest sampling unit; that is, all classrooms from a participating grade level within a selected school would participate. This recommendation was made primarily due to lack of information at the classroom level. Schools were selected with probability proportional to size (PPS) from each stratum. Within a stratum, when the number of students sampled from each school is approximately equal, the students will be selected with approximately equal probability.
- **Use of sampling weights.** Sampling weights can be applied to adjust stratum cells for under- or over-representation. In general, the use of sampling weights, when needed and appropriately assigned, can reduce bias in estimation. Another alternative is to create a self-weighted sample, in which case, every observation in the sample gets the same weight. In other words, the probability of selection is the same for every unit of observation. To achieve this, the sampling plan needs to be carefully designed. In the design, a self-weighted sample results if the following criteria are met:
  - consistent state representation in the target population and Field Test sample,
  - proportional allocation for the second-stage stratified sampling at the school grouping level,

  - under each stratum, simple random sampling (SRS) in one-stage cluster sampling if a grade within a school is the smallest sampling unit.

- **Nonresponse/nonparticipation.** The sampling needed to be designed to minimize nonresponse. A typical procedure to handle nonresponse is to act as if the characteristics of the nonrespondents within a stratum/cluster are the same as those of the respondents within the same stratum/cluster. Since a self-weighted sample with a defined sample size is intended, a replacement procedure may be implemented to adjust for nonresponse using specified replacement procedures. Using this procedure, replacement schools were selected within the same stratum to ensure that the schools declining to participate are replaced by schools with comparable characteristics (i.e., the same stratum). Alternatively, to avoid tedious replacement after sampling due to nonresponse, stratified sampling may be conducted based on the pool of Local Educational Agencies (LEAs) that indicated interest in the Field Test.

  To minimize this bias, it was of critical importance to ensure that the selected samples for replacement were representative of the Field Test populations, both in terms of performance on state-level achievement tests and demographic characteristics. Once the samples were selected with replacements, their representativeness can be evaluated using state assessment score distributions and demographic summaries comparing samples against the state-level distributions.

Detailed Sampling Procedures. The states that make up the Smarter Balanced Consortium were the primary sampling units (PSUs). PSUs generally consist of large geographic areas that are used for the sampling frame in the first stage of the multistage sample design. Stratification permits a population to be subdivided into mutually exclusive and exhaustive subpopulations. In proportionate stratified sampling, allocation of the sample is assigned to various strata that are made proportionate to the number of population elements that comprise that stratum. Within each PSU (state), additional strata were defined to increase sampling efficiency. The appropriate use of stratification can increase sample efficiency (Frankel, 1983). Stratification is most efficient when the stratum means differ widely and stratification cells are homogeneous. Within strata, homogeneity may result in significant decreases in sampling variance relative to equal-size simple random sampling. In general, it is preferable to define more strata to improve precision if the requisite background information is available and resources permit. Stratification variables were defined as ones that are related to the variable of interest, which is academic achievement.

In this complex sampling design, cluster sampling was used within strata due to administrative constraints and cost-efficiency reasons. Cluster sampling permits the selection of sample cases in groups such as schools as opposed to individuals. Although cluster sampling normally results in less information per observation than a simple random sample, its inefficiency can usually be compensated by a corresponding increase in sample size. A random sample of schools will be selected as clusters within each stratum.

A sampling frame contains the defined population necessary to implement the design, which in this case, includes students from all K-12 public schools from Smarter Balanced member states. The Quality Education Database (QED, 2012) from the MCH Corporation is a commercially available source. This database was used for sampling. One drawback is that the QED did not contain an explicit school performance variable (e.g., percent proficient). The representativeness of the resulting samples were evaluated using state demographic data. The sampling procedures for a given grade level and content area within a given Smarter Balanced member state is briefly described below.

- Step 1: For a given grade level and subject area, the number of students sampled from a given state, proportional to its size, was derived from the QED database.

- Step 2: The stratification variables used to combine schools into subgroups within each state were selected. School characteristics that are expected to relate to student performance are preferred. Ideally, state-specific achievement data, which are expected to correlate highly with performance on the Smarter Balanced assessments, was preferred. If this information was not readily available, other stratification variables of interest were considered, such as economic status (percentage of Title I students). For instance, a stratum could be defined to include schools that have a high percentage of students receiving free or reduced-price lunch. It was also necessary to limit the number of stratification variables to one or two and associated number of stratification cells from two to six in order for the sampling plan to be manageable across so many states. It was not necessary for participating states to use the exact same set of stratification variables for some subgroups, since the labels may have varied locally.
- Step 3: Classify all eligible schools within each state into two to six strata based on the stratification variable(s), and determine Field Test sample size per stratum within each state through proportional allocation. This is calculated by multiplying the number of students allocated to each state in step 1 by the percentage of students represented by the specific stratum among all strata within the state. Ideally, student population size was expected to be roughly the same across different strata.
- Step 4: For nonresponse after LEAs were initially selected for Field Test participation, a list of replacement schools was constructed. The replacement schools corresponded to a single stratum cell and were evaluated to ensure a sufficient sample was available. If not, an effort was made to recruit more schools. Selecting Field Test replacements from a list of voluntary schools avoided the potential for extensive rounds of replacements. A separate list of voluntary schools was constructed for each grade from each state.
- Step 5: Field Test participants were selected from the list of voluntary schools, if available, or from the list of all eligible schools within each stratum based on the smallest sampling unit. If the smallest sampling unit is a grade within school, a simple random sample of schools will be picked from each stratum until the overall number of students from selected schools at the grade of interest reaches or approximates the predetermined number. Multiple grade levels from participating schools might have been selected. For example, a school may be selected for grade four participation because it was also selected through stratified sampling for grade 3 participation. Some schools were selected with unique characteristics if the presence of the school in the sample was necessary to ensure sample representativeness. To ensure that Field Test candidates who were excluded from participation were not accidentally included in the sample, the selection of Field Test participants from a stratum took place after removing the excluded candidates from the eligible pool of candidates within a stratum.
- Step 6: The extent to which the selected sample is representative of the target population was evaluated for a grade and content area. Specifically, within a state the variables evaluated for representativeness might have included the following demographics:
    - performance on the last state assessment taken by the students
    - gender
    - ethnicity
    - disability
    - English-proficiency status

    - proxy for social economic status (SES)

- Step 7: Replacement schools were selected as needed. Extensive replacement was expected when the sample was selected from all eligible schools and if the state for which sampling is conducted follows the "traditional" participation model. For the particular grade of interest, the stratum to which the school that needs to be replaced belongs was identified. A school was selected from the list of all candidate schools belonging to the same stratum that has the most similar grade size consistent with the replacement school.

Sampling Results

Table 9 shows the expected sampling percentage as the target for each member state and by grade and content area for the vertical scaling. The targeted state participation was the first stage of sampling, which was intended to be proportional to state enrollment size. These results were affected to some extent by late state withdrawal from the Field Test. Due to the need to optimize the item exposure rates in test delivery, the targeted state participation rates (percent of Consortium) were not met. Tables 10 and 11 show the percentage participating for various subgroups for the Smarter Balanced Consortium (i.e., population) and the vertical scaling sample for ELA/literacy and mathematics. Overall, the student characteristics mostly matched the Smarter Balanced population characteristics. However, one of the most notable demographic differences between the target and sample was Hispanics for grade 7 mathematics. The overall number of students obtained was sufficient for conducting observed and IRT analyses. In some cases, items were not calibrated due to an insufficient number of observations per item (i.e., < 500) or score level (< 50). Tables 12, 13, and 14 show the same information for the item pool calibration sample.

Table 9. Sample Size (Percents) for ELA/literacy and Mathematics by Grade and Smarter Balanced State for Vertical Scaling.

State	Percent of Consortium	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
		ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math
California	36.4	60.5	26.6	44.9	16.8	24.5	14.2	53.9	37.0	72.4	60.7	53.8	66.2	49.9	60.7
Connecticut	3.2	2.8	18.1	8.1	28.8	17.7	24.9	7.1	15.1	2.2	3.5	6.7	2.1	11.4	6.2
Delaware	0.7	0.0	0.3	0.3	0.1	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.7	0.1	0.3
Hawaii	1.0	6.5	3.0	2.8	0.2	2.1	0.8	1.1	0.6	1.4	1.6	0.6	1.0	0.7	0.6
Idaho	1.6	1.6	5.5	3.1	8.9	6.9	11.3	2.3	5.7	0.9	2.0	2.7	2.5	15.3	16.4
Iowa	2.9	1.1	1.8	0.0	0.1	1.2	0.3	0.9	0.2	0.6	1.5	0.0	0.0	0.2	0.2
Kansas	2.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Maine	1.1	0.5	0.3	0.7	0.5	0.5	0.5	0.2	0.2	0.1	0.0	0.2	0.1	0.2	0.3
Michigan	9.2	5.5	3.0	4.3	2.7	4.4	1.9	4.1	3.9	3.6	3.0	4.2	3.8	3.6	4.9
Missouri	5.3	1.9	2.4	3.3	1.6	1.1	1.2	2.2	1.6	0.4	1.2	2.2	2.6	3.1	0.7
Montana	0.8	0.5	3.5	1.9	7.0	4.4	6.8	1.2	2.7	0.1	0.6	1.5	0.4	3.6	0.7
Nevada	2.5	3.2	2.0	2.3	1.3	2.5	1.7	2.0	2.0	0.4	2.5	2.8	2.6	1.0	1.8
New Hampshire	1.1	0.5	0.5	0.6	0.1	0.2	0.0	0.1	0.1	0.4	0.8	0.5	0.5	0.3	0.4
North Carolina	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4
North Dakota	0.6	0.3	0.2	0.2	0.3	0.2	0.2	0.2	0.1	0.1	0.0	0.2	0.3	0.1	0.1
Oregon	3.3	1.0	1.4	0.5	0.5	0.7	0.6	0.9	0.4	0.4	0.5	0.8	0.6	1.2	0.6
South Carolina	4.2	0.7	0.3	0.2	0.2	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.2
South Dakota	0.7	3.5	8.3	9.7	0.2	7.2	2.2	11.5	11.4	8.6	11.6	9.6	6.1	3.1	1.7
Vermont	0.6	0.5	0.4	0.1	0.1	0.4	0.3	0.1	0.1	0.1	0.0	0.2	0.2	0.5	0.5
US Virgin Islands	0.0	0.0	0.0	0.3	1.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
Washington	6.0	4.9	18.4	12.1	26.4	23.8	30.7	9.8	18.4	6.3	7.7	10.7	6.6	3.5	2.3
West Virginia	1.6	0.9	0.8	1.3	0.2	0.2	0.3	0.5	0.0	0.7	0.6	0.3	0.4	0.8	0.6
Wisconsin	5.0	3.7	2.7	3.0	2.0	1.9	0.8	1.0	0.4	1.2	2.0	3.0	3.1	0.0	0.0
Wyoming	0.5	0.2	0.3	0.4	0.1	0.3	0.1	0.6	0.1	0.0	0.0	0.0	0.0	0.2	0.3
Sample Size		23,223	24,799	35,689	38,925	31,594	42,380	31,535	29,946	30,913	28,271	35,913	34,880	50,657	47,608

Table 10. Student Demographic Characteristics (in Percentages) for ELA/literacy by Grade for Vertical Scaling.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample
Male	51.0	51.0	51.1	51.5	51.0	51.1	51.1	51.1	51.1	51.2	51.1	51.0	51.3	50.0
Female	48.5	49.0	48.5	48.5	48.5	48.9	48.5	48.9	48.5	48.8	48.5	49.0	48.7	50.0
Native American or Alaska Native	1.1	1.7	1.1	3.2	1.1	2.9	1.1	2.9	1.1	1.7	1.1	2.7	1.0	2.4
Asian	6.5	8.0	6.7	7.4	6.7	7.2	6.6	7.3	6.5	9.4	6.7	7.1	6.1	7.3
Native Hawaiian or other Pacific Islander	0.8	3.0	0.8	1.5	0.8	1.6	0.8	1.2	0.7	0.8	0.7	0.9	0.7	0.8
Hispanic or Latino	28.7	31.9	28.0	26.8	27.8	28.2	27.4	28.0	26.9	42.2	26.6	27.8	26.7	32.2
Black or African American	10.7	6.9	10.6	7.0	10.8	7.9	11.1	6.8	11.4	5.1	11.4	8.0	11.8	6.9
White or Caucasian	48.7	47.0	49.4	55.8	49.6	61.5	49.9	55.0	50.3	40.6	50.6	54.1	50.2	54.3
Two or More Races	3.6	4.5	3.4	4.3	3.3	4.1	3.2	4.0	3.1	2.8	3.0	3.8	2.7	3.6
Individualized Education Program	11.4	9.4	12.3	10.5	12.5	10.8	12.1	10.4	11.7	9.3	11.5	9.3	10.4	7.1
Limited English Proficient	18.0	18.8	15.3	12.6	12.6	11.0	9.8	9.7	8.7	12.4	7.8	7.7	7.1	6.0
Economic Disadvantaged	55.4	54.1	55.3	51.6	54.6	50.1	54.2	50.7	53.1	56.1	51.9	48.8	48.6	46.6

Table 11. Student Demographic Characteristics (in Percentages) for Mathematics by Grade for Vertical Scaling.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample
Male	51.0	50.5	51.1	51.3	51.0	51.0	51.1	50.8	51.1	50.5	51.1	50.4	51.3	49.6
Female	48.5	49.5	48.5	48.7	48.5	49.0	48.5	49.2	48.5	49.5	48.5	49.6	48.7	50.4
Native American or Alaska Native	1.1	3.2	1.1	4.0	1.1	3.3	1.1	2.9	1.1	2.3	1.1	1.7	1.0	1.5
Asian	6.5	7.6	6.7	6.3	6.7	6.6	6.6	6.4	6.5	9.6	6.7	10.2	6.1	9.9
Native Hawaiian or other Pacific Islander	0.8	1.2	0.8	0.9	0.8	0.9	0.8	0.9	0.7	1.4	0.7	0.6	0.7	0.7
Hispanic or Latino	28.7	30.6	28.0	26.9	27.8	23.7	27.4	23.4	26.9	35.7	26.6	38.9	26.7	37.1
Black or African American	10.7	8.7	10.6	10.6	10.8	8.5	11.1	6.9	11.4	5.7	11.4	5.5	11.8	5.4
White or Caucasian	48.7	58.5	49.4	66.0	49.6	69.0	49.9	63.6	50.3	45.2	50.6	42.6	50.2	47.6
Two or More Races	3.6	3.6	3.4	4.8	3.3	5.0	3.2	4.6	3.1	3.3	3.0	3.4	2.7	3.6
Individualized Education Program	11.4	9.7	12.3	10.8	12.5	11.0	12.1	9.4	11.7	8.7	11.5	8.2	10.4	6.7
Limited English Proficient	18.0	16.1	15.3	11.2	12.6	8.7	9.8	6.7	8.7	10.1	7.8	9.2	7.1	6.8
Economic Disadvantaged	55.4	52.3	55.3	50.8	54.6	48.1	54.2	45.3	53.1	51.6	51.9	50.8	48.6	48.7

Table 12. Sample Size (Percents) for ELA/literacy and Mathematics by Grade and Smarter Balanced State for the Item Pool Calibration.

State	Percent of Consortium	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
		ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math
California	36.4	38.6	30.0	36.1	25.8	27.5	24.2	36.1	53.9	41.6	60.2	36.7	62.1	62.7	61.5
Connecticut	3.2	14.7	17.1	15.0	21.5	19.7	21.4	17.4	11.6	15.6	9.6	18.1	8.4	12.1	12.1
Delaware	0.7	0.6	0.4	0.5	0.5	0.3	0.6	0.5	0.2	0.6	0.1	0.6	0.4	0.6	0.6
Hawaii	1.0	2.5	1.7	1.7	1.1	1.4	1.3	0.8	0.6	1.0	0.8	0.6	0.9	0.5	0.4
Idaho	1.6	4.8	7.1	5.2	9.3	6.8	10.7	5.5	2.7	5.2	1.4	5.9	1.7	7.1	8.4
Iowa	2.9	0.7	2.1	0.0	0.1	0.6	0.1	0.3	0.1	0.2	0.5	0.0	0.0	0.0	0.0
Kansas	2.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Maine	1.1	0.8	1.0	1.1	1.2	0.9	1.0	0.7	0.6	0.8	0.8	0.6	0.4	0.3	0.3
Michigan	9.2	5.6	5.0	4.9	3.9	5.5	3.5	4.2	3.0	3.7	2.6	3.8	2.8	3.4	3.8
Missouri	5.3	1.5	0.9	1.4	0.7	0.5	0.5	1.2	0.6	0.2	0.5	1.0	0.9	1.0	0.9
Montana	0.8	2.9	3.8	3.5	5.4	4.5	5.7	3.3	1.5	3.3	0.6	3.8	0.8	3.2	3.1
Nevada	2.5	2.3	2.3	2.2	1.7	2.2	1.9	2.3	1.1	1.9	1.5	2.1	1.6	1.4	1.9
New Hampshire	1.1	0.5	0.7	0.6	0.4	0.4	0.4	0.3	0.3	0.6	0.4	0.3	0.3	0.2	0.3
North Carolina	8.6	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.0	0.1	0.1
North Dakota	0.6	0.4	0.3	0.2	0.3	0.3	0.2	0.3	0.1	0.5	0.1	0.2	0.2	0.2	0.2
Oregon	3.3	1.8	2.0	1.7	1.6	1.8	1.7	1.7	1.7	1.9	1.5	1.4	1.7	0.5	0.6
South Carolina	4.2	0.5	0.2	0.3	0.2	0.5	0.2	0.2	0.0	0.2	0.0	0.0	0.0	0.4	0.4
South Dakota	0.7	4.0	4.2	6.0	2.8	5.1	4.0	5.9	3.4	5.4	3.5	5.7	2.2	3.2	3.1
Vermont	0.6	0.5	0.4	0.4	0.4	0.4	0.3	0.5	0.2	0.4	0.2	0.4	0.2	0.3	0.4
US Virgin Islands	0.0	0.0	0.0	0.5	0.8	0.0	0.0	0.0	0.0	0.7	0.7	0.0	0.0	0.1	0.1
Washington	6.0	12.0	16.1	14.1	18.8	17.5	18.4	14.7	16.7	12.6	12.7	15.4	13.3	1.7	0.9
West Virginia	1.6	0.9	0.8	0.8	0.6	0.3	1.0	1.0	0.5	0.6	0.4	0.5	0.5	0.6	0.7
Wisconsin	5.0	4.2	3.9	3.4	2.7	3.3	2.6	2.7	1.1	2.7	1.9	2.9	1.8	0.0	0.0
Wyoming	0.5	0.2	0.2	0.3	0.2	0.3	0.1	0.4	0.2	0.0	0.0	0.0	0.1	0.1	0.1
Sample Size		85,889	95,143	94,915	109,441	88,293	108,412	93,536	117,691	93,431	117,049	98,433	116,459	261,405	262,111

Table 13. Student Demographic Characteristics (in Percentages) for ELA/literacy by Grade for the Item Pool Calibration.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample
Male	51.0	51.4	51.1	51.4	51.0	51.2	51.1	51.2	51.1	51.4	51.1	51.6	51.3	51.2
Female	48.5	48.6	48.5	48.6	48.5	48.8	48.5	48.8	48.5	48.6	48.5	48.4	48.7	48.8
Native American or Alaska Native	1.1	2.8	1.1	3.2	1.1	3.1	1.1	3.3	1.1	3.2	1.1	3.2	1.0	1.9
Asian	6.5	7.3	6.7	7.5	6.7	7.1	6.6	6.8	6.5	7.6	6.7	6.9	6.1	8.2
Native Hawaiian or other Pacific Islander	0.8	1.5	0.8	1.1	0.8	1.2	0.8	0.9	0.7	0.8	0.7	0.9	0.7	0.9
Hispanic or Latino	28.7	30.2	28.0	28.4	27.8	28.6	27.4	28.8	26.9	32.8	26.6	27.8	26.7	30.3
Black or African American	10.7	10.0	10.6	9.3	10.8	10.2	11.1	10.4	11.4	9.9	11.4	10.6	11.8	9.9
White or Caucasian	48.7	54.1	49.4	56.6	49.6	58.8	49.9	57.3	50.3	52.3	50.6	57.3	50.2	50.3
Two or More Races	3.6	4.1	3.4	4.2	3.3	4.0	3.2	3.9	3.1	3.5	3.0	3.4	2.7	3.2
Individualized Education Program	11.4	10.3	12.3	10.9	12.5	11.5	12.1	11.1	11.7	10.4	11.5	10.4	10.4	8.1
Limited English Proficient	18.0	16.6	15.3	13.6	12.6	11.1	9.8	9.7	8.7	9.7	7.8	7.4	7.1	6.2
Economic Disadvantaged	55.4	53.4	55.3	51.9	54.6	50.6	54.2	51.1	53.1	52.8	51.9	48.6	48.6	46.2

Table 14. Student Demographic Characteristics (in Percentages) for Mathematics by Grade for the Item Pool Calibration.

Demographic Group	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		High School	
	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample
Male	51.0	51.2	51.1	51.3	51.0	51.4	51.1	51.1	51.1	51.5	51.1	51.2	51.3	50.9
Female	48.5	48.8	48.5	48.7	48.5	48.6	48.5	48.9	48.5	48.5	48.5	48.8	48.7	49.1
Native American or Alaska Native	1.1	3.0	1.1	4.4	1.1	4.3	1.1	2.1	1.1	1.9	1.1	2.0	1.0	2.0
Asian	6.5	7.0	6.7	7.0	6.7	6.9	6.6	8.2	6.5	8.3	6.7	8.2	6.1	8.5
Native Hawaiian or other Pacific Islander	0.8	1.4	0.8	1.1	0.8	1.1	0.8	0.9	0.7	1.0	0.7	0.9	0.7	0.9
Hispanic or Latino	28.7	31.3	28.0	29.0	27.8	27.7	27.4	33.7	26.9	36.2	26.6	33.0	26.7	31.5
Black or African American	10.7	9.8	10.6	9.7	10.8	8.5	11.1	7.9	11.4	8.3	11.4	8.0	11.8	9.9
White or Caucasian	48.7	56.2	49.4	59.3	49.6	61.4	49.9	50.7	50.3	46.9	50.6	49.7	50.2	49.0
Two or More Races	3.6	3.8	3.4	4.4	3.3	4.4	3.2	4.4	3.1	3.7	3.0	3.9	2.7	3.2
Individualized Education Program	11.4	10.4	12.3	11.1	12.5	11.3	12.1	11.1	11.7	10.5	11.5	10.2	10.4	8.0
Limited English Proficient	18.0	17.6	15.3	13.3	12.6	10.5	9.8	11.5	8.7	11.0	7.8	9.4	7.1	7.2
Economic Disadvantaged	55.4	53.1	55.3	52.1	54.6	50.0	54.2	52.2	53.1	53.3	51.9	49.9	48.6	46.4

Field Test Administration and Security

Student Participation. All students in the specified grade levels were eligible to participate in the Smarter Balanced Field Test unless they received a special exemption. In general, if a student participated in the Consortium state's general education accountability assessment or took the Alternate Assessment based on Modified Achievement Standards (AA-MAS) and attended a school participating in the Field Test, the student was eligible to participate. Consistent with the Smarter Balanced field-testing plan, all students, including students with disabilities, English language learners (ELLs), and ELLs with disabilities, had an equal opportunity to participate in the Smarter Balanced Field Test. All students enrolled in grades 3–11 selected to participate in the Smarter Balanced ELA/literacy or mathematics assessment are required to participate except

- students with the most significant cognitive disabilities who meet the criteria for the mathematics alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population);
- students with the most significant cognitive disabilities who meet the criteria for the English language/literacy alternate assessment based on alternate achievement standards (approximately 1% or fewer of the student population); and
- ELLs who enrolled within the last 12 months prior to the beginning of testing in a US school and have a one-time exemption. These students may instead participate in their state's English language proficiency assessment consistent with state and federal policy.

Practice and Training Tests. To expose students to various items types and other features of the Field Test in ELA/literacy and mathematics, it was highly recommended that all students complete the Practice Test and/or the Training Test for the Field Test. Each resource offered students a unique opportunity to experience the testing situation in a manner similar to what was experienced on the Field Test. Practice tests were grade-specific (3–8 and 11) and included a range of item types, grade-level content, and difficulty. There were approximately 30 items on a Practice Test in each content area for the Field Test. In addition, the Practice Tests included an initial set of accessibility features that were available to all students such as highlighting text, zooming in and out, marking items for review, and the digital notepad. A user guide provided direct guidance on accessing the Practice Tests, as well as frequently asked questions. The Training Tests were not grade specific and provided students and teachers with an opportunity to become familiar with the software and all interface features and functionalities that were used in the Smarter Balanced Field Test. They were available by grade bands (3–5, 6–8, and high school) and had six items in ELA/literacy and eight to nine in mathematics. The Training Tests did not include performance tasks. Table 15 summarizes the features of the Training and Practice Tests.

Table 15. Comparison of Features for the Training and Practice Tests.

Feature	Practice Test	Training Test
Purpose	Provide students the opportunity to experience a range of grade-specific item types (as well as performance tasks) similar in format and structure to the Smarter Balanced assessments.	Provide students with an opportunity to become familiar with the software and interface features used in the Smarter Balanced assessments.
Grade Level	Individual assessments at each grade • Grades 3–8 and 11	Three assessments by grade band: • Grades 3–5 • Grades 6–8 • High School
Type of Items	Approximately 30 items in ELA/literacy and 30 items in mathematics per grade level One ELA/literacy and one mathematics performance task available per grade level	Approximately 14–15 items per grade band (6 in ELA/literacy and 8–9 in mathematics) No performance tasks Included new item types not present in the practice test (matching tables, table fill-in, & evidence-based selected response)
Available Embedded Universal Tools, Designated Supports, and Accommodations	All universal tools Most designated supports, including: • Color contrast • Masking • Text-to-speech items • Translations (glossary): Spanish Most accommodations, including: • American Sign Language for all mathematics items and ELA/literacy listening stimuli and items • Braille • Streamlining	All universal tools All designated supports, including: • Color contrast • Masking • Text-to-speech items • Translated test directions: Spanish • Translations (glossary): Spanish, Arabic, Cantonese, Filipino, Korean, Mandarin, Punjabi, Russian, Ukrainian, Vietnamese • English glossary • Full translation: Spanish All accommodations, including: • American Sign Language for all mathematics items and ELA/literacy listening stimuli and items • Braille • Streamlining • Text-to-speech for reading passages in grades 6 to high school

General Field Test Administration Procedures. A brief overview of the general test administration rules are provided as well as information about various test tools and accommodations.

- CAT items (i.e., non-performance tasks) and performance tasks were presented in the Field Test administration as separate tests. All students participating in the Field Test, regardless of content area (ELA/literacy or mathematics) received CAT items, a classroom activity, and a performance task. In some cases, schools choose to administer both content areas to either the same or the different groups of students.
- The number of items in the CAT portion of the Field Test varied.
- The tests were not timed, so all time estimates were approximate. Students were allowed extra time if needed.
- The Field Test could be spread out over multiple days as needed.
- The Classroom Activity had to be completed prior to administration of the performance task.
- Students were not permitted to return to a test once it had been completed and submitted.
- Within each test, there may be several segments. A student was not permitted to return to a segment once it had been completed and submitted as complete.
- Students were instructed to answer all test items on a page before going to the next one. Some pages (i.e., screens) contained multiple test items. Students used a vertical scroll bar to view all items on a page.
- Students were required to answer all test items before submission for final processing.
- Students could mark items for review and use the Past/Marked drop-down list to return to those items.

The recommended order for test administration was to implement the CAT followed by the performance task assessment. For the performance task, the Classroom Activity was conducted, followed by the individually administered, online performance task. The recommendation was to administer the performance task portion of the assessment on a separate day from the CAT. For the performance tasks, an additional recommendation was that students might be best served by sequential, uninterrupted time that may exceed the time allotted in a student's regular classroom schedule.

During the CAT portion of the test, if a test was paused for more than 20 minutes the student was

- required to log back into the student interface;
- presented with the test page containing the test item(s) he or she was working on when the test was paused (if the page contains at least one unanswered item) or with the next test page (if all items on the previous test page were all answered); and,
- not permitted to review or change any previously answered items (with the exception of items on a page that contains at least one item that was not answered yet).

During the performance task portion of the test, there were no pause restrictions. If a test was paused for 20 minutes or more, the student could return to the section and continue typing his or her responses. Any highlighted text, notes on the digital notepad, or items marked for review were not saved when a test was paused. In the event of a technical issue (e.g., power outage or network failure), students were logged out and the test was automatically paused. Students needed to log in again when resuming the test.

As a security measure, students were automatically logged out of the test after 20 minutes of test inactivity. Activity was defined as selecting an answer or navigation option in the test (e.g., clicking [Next] or [Back] or using the Past/Marked Questions drop-down list to navigate to another item). Moving the mouse or clicking on an empty space on the screen was not considered test-taking activity. Before the system logged the student out of the test, a warning message was displayed on the screen. If the student did not click [Ok] within 30 seconds after the message appeared, he or she was logged out. Clicking [Ok] restarted the 20-minute inactivity timer.

A student's CAT administration remained active until the student completed and submitted the test or 45 calendar days elapsed after the student had initiated testing, whichever occurred sooner. A second recommendation was to minimize the amount of time between beginning and completing each test within a content area. Smarter Balanced suggest that students complete the CAT portion of the test within five days of starting the designated content area. The performance task was a separate test that remained active only for ten calendar days after the student began the performance task. However, Smarter Balanced recommended that students complete the PT within three days of starting.

Test Windows, and Testing Time. The Field Test was administered March 18–June 6, 2014. For the Field Test, schools were asked to select an anticipated testing window or were provided a testing window by their state. Smarter Balanced used this information to ensure that there was sufficient server capacity for all scheduled students to test.

Table 16 contains the estimated time required for most students to complete the Smarter Balanced Field Test based on the Pilot Test. Classroom Activities were designed to fit into a 30-minute window and will vary due to the complexity of the topic and individual student needs. These estimates did not account for any time needed to start computers, load secure browsers, and log in students. Note that the duration, timing, break/pause rules, and session recommendations varied in each content area and component.

Table 16. Expected Testing Times for Smarter Balanced Field Tests.

Content Area	Grades	CAT	Performance Task	Total	Classroom Activity	Overall Total
ELA/literacy	3–5	1: 30	2:00	3:30	0:30	4:00
	6–8	1:30	2:00	3:30	0: 30	4:00
	HS	2:00	2:00	4:00	0: 30	4:30
Mathematics	3–5	1:30	1:00	2:30	0: 30	3:00
	6–8	2:00	1:00	3:00	0: 30	3:30
	HS	2:00	1:30	3:30	0: 30	4:00
Combined	3–5	3:00	3:00	6:00	1:00	7:00
	6–8	3:30	3:00	6:30	1:00	7:30
	HS	4:00	3:30	7:30	1:00	8:30

Test Duration (Testing Time)

The Smarter Balanced tests were untimed. For test administration planning purposes, some indication of testing time is necessary. The delivery system was not able to give a per item student response time that could be accumulated accurately corresponding to test time for a student. A rough estimate was constructed that corresponded to test duration. Test duration was defined here as when the student entered the administration until the "submit" button was pressed that ended the assessment component. Since tests are administered as separate components, test duration is computed for ELA/literacy and mathematics and for both CAT and performance tasks. The resulting test durations are shown in Tables 17 to 20, which show the number of students by duration range and the corresponding cumulative percentage. Test duration is given in minutes. For ELA/literacy, most students required more than 90 minutes to complete the CAT and performance task components. The mathematics performance tasks were for the most part completed in 90 minutes. Note that longer test duration could result from test sessions that occurred over several days.

Table 17. Distribution of Test Duration in Minutes for the ELA/literacy CAT for the Item Pool Calibration Administration.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
>= 90	49,830	100.0	58,815	100.0	58,502	100.0	62,155	100.0	54,138	100.0	56,252	100.0	71,053	100.0
85 - 90	1,729	40.7	2,156	37.1	2,117	32.6	2,393	32.3	2,528	40.2	2,587	41.0	6,094	69.6
80 - 85	1,913	38.7	2,305	34.8	2,279	30.2	2,770	29.6	2,722	37.4	2,970	38.2	7,160	67.0
75 - 80	2,195	36.4	2,547	32.3	2,352	27.5	2,902	26.6	3,162	34.4	3,223	35.1	8,766	64.0
70 - 75	2,449	33.8	2,724	29.6	2,534	24.8	3,074	23.5	3,357	30.9	3,452	31.7	10,187	60.2
65 - 70	2,782	30.9	2,843	26.7	2,638	21.9	3,047	20.1	3,694	27.2	3,642	28.1	12,059	55.9
60 - 65	3,069	27.6	3,082	23.7	2,794	18.9	3,097	16.8	3,793	23.1	3,815	24.3	13,574	50.7
55 - 60	3,198	23.9	3,309	20.4	2,634	15.6	2,861	13.4	3,575	18.9	3,746	20.3	14,645	44.9
50 - 55	3,355	20.1	3,275	16.8	2,663	12.6	2,483	10.3	3,401	15.0	3,456	16.4	15,254	38.6
45 - 50	3,291	16.1	3,260	13.3	2,393	9.5	2,136	7.6	2,877	11.2	3,181	12.7	15,270	32.1
40 - 45	3,208	12.2	3,010	9.8	2,077	6.8	1,762	5.3	2,438	8.1	2,765	9.4	14,582	25.6
35 - 40	2,686	8.4	2,509	6.6	1,613	4.4	1,222	3.3	1,885	5.4	2,297	6.5	13,049	19.4
30 - 35	2,015	5.2	1,743	3.9	1,041	2.5	831	2.0	1,296	3.3	1,644	4.1	10,836	13.8
25 - 30	1,226	2.8	1,061	2.1	633	1.3	498	1.1	824	1.8	1,144	2.4	8,156	9.2

Table 17. Distribution of Test Duration in Minutes for the ELA/literacy CAT for the Item Pool Calibration Administration continued.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
20 - 25	672	1.3	556	0.9	331	0.6	316	0.6	479	0.9	623	1.2	6,162	5.7
15 - 20	351	0.5	250	0.3	149	0.2	157	0.2	278	0.4	346	0.5	4,519	3.0
10 - 15	81	0.1	64	0.1	33	0.0	31	0.0	75	0.1	93	0.1	1,941	1.1
5 - 10	9	0.0	13	0.0	7.0	0.0	11	0.0	14	0.0	34	0.0	636	0.3

Table 18. Distribution of Test Duration in Minutes for the ELA/literacy Performance Task for the Item Pool Calibration.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
>= 90	45,051	100.0	54,619	100.0	49,075	100.0	48,593	100.0	42,699	100.0	48,834	100.0	41,013	100.0
85 - 90	1,170	44.8	1,680	40.5	1,693	42.2	2,034	45.8	1,752	51.4	2,039	47.4	3,479	75.2
80 - 85	1,513	43.4	1,880	38.6	1,967	40.2	2,235	43.5	2,080	49.4	2,360	45.2	4,120	73.1
75 - 80	1,697	41.6	2,151	36.6	2,196	37.9	2,448	41.0	2,392	47.1	2,491	42.7	4,816	70.6
70 - 75	1,937	39.5	2,343	34.3	2,458	35.3	2,760	38.3	2,700	44.3	2,764	40.0	5,613	67.7
65 - 70	2,180	37.1	2,603	31.7	2,539	32.4	2,994	35.2	2,903	41.3	3,082	37.0	6,347	64.3

Table 18. Distribution of Test Duration in Minutes for the ELA/literacy Performance Task for the Item Pool Calibration continued.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
60 - 65	2,479	34.4	2,759	28.9	2,718	29.4	3,261	31.9	3,294	38.0	3,249	33.7	7,170	60.4
55 - 60	2,679	31.4	3,069	25.9	2,794	26.2	3,397	28.2	3,463	34.2	3,335	30.2	7,843	56.1
50 - 55	3,149	28.1	3,243	22.5	3,104	22.9	3,359	24.4	3,620	30.3	3,571	26.6	9,031	51.4
45 - 50	3,284	24.3	3,339	19.0	3,253	19.3	3,338	20.7	3,822	26.2	3,726	22.8	10,008	45.9
40 - 45	3,481	20.2	3,255	15.3	3,075	15.4	3,364	17.0	4,010	21.8	3,722	18.8	11,008	39.8
35 - 40	3,427	16.0	2,981	11.8	2,834	11.8	3,169	13.2	3,890	17.3	3,529	14.8	11,457	33.2
30 - 35	3,099	11.8	2,588	8.5	2,404	8.5	2,893	9.7	3,591	12.8	3,207	11.0	11,508	26.3
25 - 30	2,702	8.0	2,201	5.7	2,079	5.6	2,456	6.5	3,289	8.7	2,920	7.5	11,400	19.3
20 - 25	2,192	4.7	1,737	3.3	1,589	3.2	1,909	3.7	2,501	5.0	2,304	4.4	10,893	12.4
15 - 20	1,631	2.0	1,315	1.4	1,108	1.3	1,427	1.6	1,900	2.2	1,775	1.9	9,608	5.8

Table 19. Distribution of Test Duration in Minutes for the Mathematics CAT for the Item Pool Calibration.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
>= 90	24,813	100.0	36,112	100.0	43,479	100.0	34,510	100.0	37,517	100.0	38,256	100.0	37,933	100.0
85 - 90	1,085	73.2	1,984	66.5	2,367	59.5	2,180	70.1	1,575	66.4	1,504	65.5	2,249	83.1
80 - 85	1,423	72.0	2,577	64.7	3,021	57.2	2,495	68.2	1,869	65.0	1,883	64.1	2,919	82.1
75 - 80	1,784	70.5	2,887	62.3	3,537	54.4	3,152	66.1	2,429	63.3	2,363	62.4	3,867	80.8
70 - 75	2,292	68.5	3,549	59.6	4,047	51.1	3,938	63.4	2,887	61.2	2,983	60.3	5,009	79.1
65 - 70	2,936	66.1	4,291	56.3	4,808	47.4	4,811	59.9	3,526	58.6	3,564	57.6	6,225	76.9
60 - 65	3,612	62.9	5,233	52.3	5,461	42.9	5,636	55.8	4,121	55.4	4,328	54.4	7,917	74.1
55 - 60	4,547	59.0	6,100	47.5	6,171	37.8	6,469	50.9	5,062	51.7	4,891	50.5	9,887	70.6
50 - 55	5,632	54.1	7,203	41.8	6,621	32	7,432	45.3	5,808	47.2	5,746	46.1	12,076	66.2
45 - 50	6,924	48.0	7,888	35.1	7,049	25.8	8,103	38.9	6,958	42.0	6,513	40.9	15,206	60.8
40 - 45	7,849	40.5	8,112	27.8	6,540	19.3	8,356	31.9	7,739	35.8	7,224	35.0	18,451	54.0
35 - 40	8,437	32.0	7,678	20.3	5,679	13.2	8,517	24.6	8,366	28.8	8,269	28.5	21,056	45.8
30 - 35	8,041	22.9	6,046	13.2	4,016	7.9	7,526	17.2	8,001	21.4	7,905	21.0	22,147	36.4
25 - 30	6,516	14.2	4,332	7.5	2,476	4.1	5,906	10.7	6,920	14.2	6,511	13.9	20,762	26.6

Table 19. Distribution of Test Duration in Minutes for the Mathematics CAT for the Item Pool Calibration continued.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
20 - 25	4,274	7.2	2,553	3.5	1,297	1.8	3,736	5.6	4,743	8.0	4,670	8.0	17,463	17.3
15 - 20	1,975	2.6	1,026	1.2	555	0.6	1,882	2.4	2,722	3.7	2,730	3.8	12,992	9.6
10 - 15	371	0.4	189	0.2	86.0	0.1	745	0.8	1,199	1.3	1,163	1.3	6,583	3.8
5 - 10	28	0.0	36	0.0	16	0.0	125	0.1	265	0.2	300	0.3	1,944	0.9

Table 20. Distribution of Test Duration in Minutes for the Mathematics Performance Tasks.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
>= 90	5,514	100.0	7,396	100.0	11,447	100.0	9,067	100.0	7,722	100.0	8,935	100.0	9,188	100.0
85 - 90	452	93.9	704	93.0	1,030	89.0	671	90.3	207	91.5	302	90.4	558	94.5
80 - 85	571	93.4	896	92.3	1,159	88.0	876	89.5	287	91.3	472	90.1	733	94.1
75 - 80	779	92.7	1,121	91.5	1,626	86.9	1,135	88.6	393	91.0	625	89.6	1,031	93.7
70 - 75	991	91.9	1,481	90.4	2,152	85.3	1,346	87.4	473	90.5	902	88.9	1,408	93.1
65 - 70	1,227	90.8	1,941	89.0	2,730	83.2	1,799	85.9	676	90.0	1,142	87.9	1,888	92.2
60 - 65	1,787	89.4	2,645	87.2	3,548	80.6	2,317	84.0	917	89.3	1,597	86.7	2,557	91.1

Table 20. Distribution of Test Duration in Minutes for the Mathematics Performance Tasks continued.

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		HS	
Range	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent	No.	Percent
55 - 60	2,585	87.4	3,375	84.7	4,634	77.2	3,252	81.5	1,253	88.3	2,018	85.0	3,663	89.5
50 - 55	3,329	84.6	4,581	81.5	5,820	72.7	4,147	78.0	1,817	86.9	2,597	82.8	4,962	87.3
45 - 50	4,550	80.9	6,323	77.2	7,340	67.1	5,412	73.6	2,766	84.9	3,799	80.0	7,091	84.4
40 - 45	6,014	75.8	8,196	71.2	8,844	60.0	7,254	67.8	4,347	81.9	5,306	75.9	9,847	80.1
35 - 40	8,118	69.1	10,273	63.5	10,675	51.5	9,134	60.0	6,752	77.1	7,414	70.2	13,388	74.2
30 - 35	10,214	60.1	12,780	53.8	11,642	41.3	11,036	50.2	9,837	69.7	9,942	62.2	17,129	66.1
25 - 30	12,095	48.8	14,615	41.7	11,900	30.0	11,840	38.3	13,081	58.9	12,189	51.6	21,009	55.8
20 - 25	13,288	35.4	14,041	27.9	10,098	18.6	11,464	25.6	15,711	44.5	13,651	38.5	24,165	43.1
15 - 20	11,387	20.6	10,484	14.6	6,359	8.9	8,095	13.3	15,031	27.2	13,025	23.8	25,152	28.6
10 - 15	7,183	8.0	5,011	4.7	2,838	2.7	4,312	4.6	9,767	10.7	9,083	9.8	22,352	13.5

Universal Tools, Designated Supports, and Accommodations

The Smarter Balanced Assessment Consortium's *Usability, Accessibility, and Accommodations Guidelines* are intended for school-level personnel and decision-making teams, including Individualized Education Program (IEP) teams, as they prepare for and implement the Smarter Balanced assessments. The Guidelines provide information for classroom teachers, English development educators, special education teachers, and related services personnel to use in selecting and administering universal tools, designated supports, and accommodations for those students requiring them. The Guidelines are also intended for assessment staff and administrators who oversee the decisions that are made in instruction and assessment.

The Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* apply to all students. They emphasize an individualized approach to the implementation of assessment practices for those students who have diverse needs and participate in large-scale content assessments. The Guidelines focus on universal tools, designated supports, and accommodations for the Smarter Balanced content assessments of English language arts/literacy and mathematics. At the same time, the Guidelines support important instructional decisions about accessibility and accommodations for students who participate in the Smarter Balanced assessments. The Guidelines recognize the critical connection between accessibility and accommodations in instruction and accessibility and accommodations during assessment. The Field Test and Training Tests contained embedded universal tools, designated supports, and accommodations and are defined in Table 21. Embedded resources are those that are part of the computer administration system, whereas non-embedded resources are provided outside of that system. Chapter 5 on Test Fairness presents a more comprehensive discussion of these issues.

Table 21. Definitions for Universal Tools, Designated Supports, and Accommodations.

Type	Definition
Universal Tools	Access features of the assessment that either are provided as digitally delivered components of the test administration system or separate from it. Universal tools are available to all students based on student preference and selection.
Designated Supports	Access features of the assessment available for use by any student for whom the need has been indicated by an educator (or team of educators working with the parent/guardian and student). They either are provided as digitally delivered components of the test administration system or separate from it.
Accommodations	Accommodations are changes in procedures or materials that increase equitable access during the Smarter Balanced assessments. Assessment accommodations generate valid assessment results for students who need them; they allow these students to show what they know and can do. Accommodations are available for students with documented IEPs or Section 504 Plans. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standards, or intended outcome of the assessment.

Test Security

The test environment refers to all aspects of the testing situation. The test environment includes what a student can see, hear, or access (including access via technology). Requirements of a secure test environment include, but are not limited to, the following:

- Providing a quiet environment, void of talking or other distractions that might interfere with a student's ability to concentrate or might compromise the testing situation.
- Actively supervising students to prevent access to unauthorized electronic devices that link to outside information, communication among students, and photographing or otherwise copying test content.
- Removing information displayed on bulletin boards, chalkboards or dry-erase boards, or charts (e.g., wall charts that contain literary definitions, maps, mathematics formulas, etc.) that might assist students in answering questions must be removed.
- Seating students so there is enough space between them to minimize opportunities to view each other's work, or providing students with tabletop partitions.
- Allowing students access to only the allowable resources identified by Smarter Balanced specific to the assessment (or that portion of an assessment).
- Allowing only students who are testing to observe assessment items. Students who are not being assessed or unauthorized staff should be removed from the testing environment.
- Administering the Smarter Balanced Field Test only through the Student Interface via a secure browser.

Item security rules included, but were not limited to, the following:

- Unless assigned as an accommodation, no copies of the test items, stimuli, reading passages, performance task materials, or writing prompts could be made or otherwise retained. This rule included any digital, electronic, or manual device used to record or retain item information.
- Descriptions of test items, stimuli, printed reading passages, or writing prompts must not be retained, discussed, or released to anyone. All printed test items, stimuli, and reading passages must be securely shredded immediately following a test session.
- Test items, stimuli, reading passages, or writing prompts must never be sent by e-mail or fax or replicated/displayed electronically.
- Secure test items, stimuli, reading passages, or writing prompts must not be used for instruction.
- No review, discussion, or analysis of test items, stimuli, reading passages, or writing prompts were allowed at any time by students, staff, or teaching assistants, including before, during, or between sections of the test. Student interaction with test content during a test was limited to what was dictated for the purpose of a performance task that was standardized.
- No form or type of answer key may be developed for test items.

Test security incidents, such as improprieties, irregularities, and breaches, were behaviors prohibited during test administration, either because they lent a student a potentially unfair advantage or because they compromised the secure administration of the assessment. Whether intentional or by accident, failure to comply with security rules, either by staff or students, constituted a test security

incident. Improprieties, irregularities, and breaches were reported in accordance with each severity level. Definitions of three types of test security incidents are given in Table 22.

Table 22. Definitions for Three Levels of Test Security Incidents.

Type	Definition
Impropriety	An unusual circumstance that has a low impact on the individual or group which has a low risk of potentially affecting student performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level. An example of an impropriety might include posting a practice item to a social media site by a student.
Irregularity	An unusual circumstance that affects an individual or group of students who are testing and may potentially influence student performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level, but submitted in the online system for resolution of the appeal for testing impact.
Breach	An event that poses a threat to the validity of the test. These circumstances have external implications for the Consortium and may result in a decision to remove the test item(s) from the available secure bank. A breach incident must be reported immediately.

Test monitors were instructed to be vigilant before, during, and after testing for any situations that could lead to or be an impropriety, irregularity, or breach. The following instructions were given:

- Actively supervise students throughout the test session to ensure that students do not access unauthorized electronic devices, such as cell phones, or other unauthorized resources or tools at any time during testing.
- Make sure students clear their desks of and put away all books, backpacks, purses, cell phones, electronic devices of any kind, as well as other materials not explicitly permitted for the test.
- Make sure the physical conditions in the testing room meet the criteria for a secure test environment. Students should be seated so there is enough space between them to minimize opportunities to view another student's work.
- Students who are not being tested and unauthorized staff must not be in the room where a test is being administered. Determine where to send these students during testing and prepare appropriate assignments for them as needed.
- Make sure no instructional materials directly related to the content of the tests are visible to students, including posters or wall charts.
- States should ensure that specific guidance is provided for districts that have minimal personnel and may experience potential conflicts of interest in the identification, investigation, and/or reporting of test security incidents.

References

Folk, V. G. & Smith, R. L. (2002). Models for delivery of CBTs. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.). *Computer-based Testing: Building the Foundation for Future Assessments* (pp. 41-66). Mahwah, New Jersey: Lawrence Erlbaum.

Frankel, M. (1983). Sampling theory. In *Handbook of Survey Research*, Wright, Anderson, & Rossi (Eds.). New York: Academic Press.

Gibson, W. M. & Weiner, J. A. (1998). Generating Random Parallel Test Forms Using CTT in a Computer-based Environment. *Journal of Educational Measurement*, 35, 297-310.

Hetter, R. D. & Sympson, J. B., (1997). Item Exposure Control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.). *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, DC: American Psychological Association.

Khatri, N., Reve, A. L., & Kane, M. B. (1998). *Principles and Practices of Performance Assessment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Quality Education Data. School Year 2011-2012. MCH. Sweet Springs: MO.